Exercise testing scores as an example of better decisions through science

EUAN ASHLEY, JONATHAN MYERS, and VICTOR FROELICHER

Cardiology Division, Veterans Affairs Palo Alto Health Care System, Stanford University, Palo Alto, CA and Radcliff Infirmary, Oxford University, Oxford UNITED KINGDOM

ABSTRACT

ASHLEY, E., J. MYERS, and V. FROELICHER. Exercise testing scores as an example of better decisions through science. *Med. Sci. Sports Exerc.*, Vol. 34, No. 8, pp. 1391–1398, 2002. **Introduction:** The application of common statistical techniques to clinical and exercise test data has the potential to become a useful tool for assisting in the diagnosis of coronary artery disease, assessing prognosis, and reducing the cost of evaluating patients with suspected coronary disease. Since general practitioners function as gatekeepers and decide which patients must be referred to the cardiologist, they need to optimally use the basic tools they have available (i.e., history, physical exam, and the exercise test). **Methods:** Review of the literature with a focus on the scientific techniques for aiding the decision-making process. **Results:** Scores derived from multivariable statistical techniques considering clinical and exercise data have demonstrated superior discriminating power when compared using receiver-operating-characteristic curves with the ST segment response. In addition, by stratifying patients as to probability of disease and prognosis, they provide a management strategy. While computers as part of information management systems can calculate complicated equations to provide scores, physicians are reluctant to trust them. Thus, these scores have been represented as nomograms or simple additive tables so physicians are comfortable with their application. Scores have also been compared with physician judgment and been found to estimate the presence of coronary disease and prognosis as well as expert cardiologists, and often better than nonspecialists. **Conclusion:** Multivariate scores can empower the clinician to assure the cardiac patient with access to appropriate and cost-effective cardiological care. **Key Words:** CORONARY ARTERY DISEASE, DIAGNOSIS, BIOSTATISTICS, CLINICAL MEDICINE

INTRODUCTION

Coronary artery disease (CAD) continues to be the leading cause of morbidity and mortality in the United States. CAD prevalence is expected to remain high given the increasing proportion of the population that is elderly (8,39,41). In spite of efforts to control costs, health care costs had the greatest increase in 1999 relative to the decade. With half of the cost increase due to pharmaceuticals that can decrease heart disease interventions and events, the next target of cost containment must be expensive diagnostics and interventions. It is important to implement clinically cost-effective strategies that direct the appropriate patients to the optimal procedures. There is a growing awareness of the need to apply statistical techniques to develop evidence-based scores for better decision-making (43). The purpose of this review is to discuss these techniques in the context of the diagnostic and prognostic performance of the exercise test.

Criteria for Evaluating Diagnostic Techniques

Studies describing the value of diagnostic techniques must be evaluated by standardized rules to determine their validity. Biostatisticians have presented these rules so that diagnostic technologies can properly be evaluated before they are adopted for practice (16,30,34,35,40). Common mistakes made in studies evaluating diagnostic tests based on these rules are presented in Table 1. Critical to fulfilling the rules are that only consecutive nondiagnosed patients presenting with the symptoms or signs of the disease being diagnosed are used to evaluate the test or score and that work up-bias be reduced (11). These basic precepts have often not been followed even in studies of newer technologies than exercise testing. When these rules are not followed, the estimates of test characteristics from the studies can be erroneous.

STATISTICAL TECHNIQUES

When developing a prediction rule, investigators consider variables that they believe may predict the occurrence of the outcome. The variables found to have discriminating power (consisting of clinical information and treadmill responses) are combined to form an algorithm for estimating the probability of coronary artery disease. Many mathematical techniques are available for demonstrating what variables are predictive as well as their relative predictive

TABLE 1. The common mistakes made by researchers attempting to determine the diagnostic characteristics of a test.

- Using a target population that only consists of normal subjects and those with severe disease (Limited Challenge)
- Failure to limit work up bias
- Using heart rate targets to exclude patients
- Inclusion of MI patients
- Use of surrogates instead of appropriate measurements or endpoints

0195-9131/02/3408-1391/$3.00/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

Submitted for publication August 2001.
Accepted for publication January 2002.

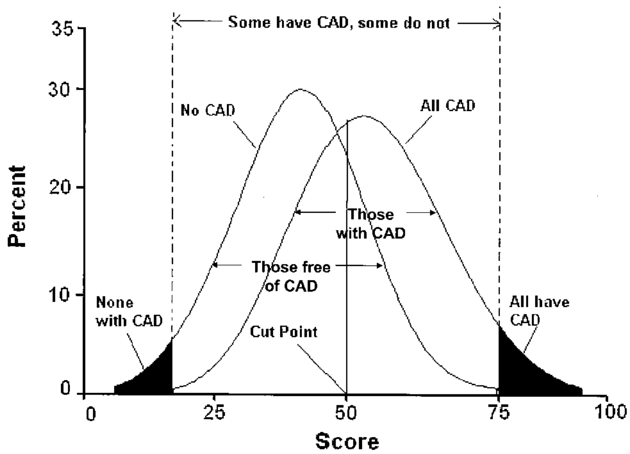

FIGURE 1—Range of characteristics plots for the simple treadmill score for those with and those without angiographic coronary disease.

power. Regression analysis methods are especially attractive because they make it possible to derive complex regression functions directly from a database. Logistic regression has been preferred since it models the relationship to a sigmoid curve (which often is the mathematical relationship between a prediction variable and an outcome) and its output is between zero and one, representing the probability of disease being present (i.e., from 0 to 100% probability of the predicted outcome) (22).

The ability of any score or measurement to diagnose a disease (e.g., CAD) depends upon how much the score differs among those with and without the disease. These measurements could be ST segment depression, calcium score using Electron Beam Computed Tomography, perfusion scan values, or echocardiographic wall motion estimates. Figure 1 consists of actual data from over one thousand male veterans who underwent both exercise testing and coronary angiography. Unfortunately, as illustrated in the figure, the values for the score or measurement usually greatly overlap. The better the test or measurement, the further apart the curves, and the less they overlap. The cut-point of 50 that we chose is a practical choice for the treadmill score we use so that those above 50 are considered to have disease and those below are considered free of CAD. However, as can be seen, this is not really the case. Figure 2 shows the trade-off between sensitivity and specificity.

Score Evaluation (ROC Curves)

The accuracy of the model to separate disease from nondisease is assessed by means of the area under a receiver-operating-characteristic (ROC) curve. ROC curve analysis is based on the plotting of sensitivity and specificity for a range of cut-points (criteria for abnormal) for a test measurement or the value of a score. The area ranges from 0 to 1, with 0.5 corresponding to no discrimination (i.e., random performance), 1.0 to perfect discrimination, and values less than 0.5 to worse-than-random performance. Most prediction rules, like other diagnostic measures, have a range of possible results. Several possible cutoff criteria could be used to separate results into positive and negative groups. For each criterion chosen, the rule will have a different sensitivity and specificity. An ROC curve is a plot of the sensitivity versus specificity for the full range of the score. The shape of the curve shows the trade-offs between sensitivity and specificity produced at different criteria with specificity and sensitivity being inversely related (36). Figure 3 presents an ROC plot of our simple treadmill score ranging from 0 to 100%. Other cut-points, such as 40 and 60, could be appropriate for particular purposes of the test, such as screening healthy people where a high specificity is needed, or for ruling out ischemia after presentation to an Emergency Department with chest pain where high sensi-

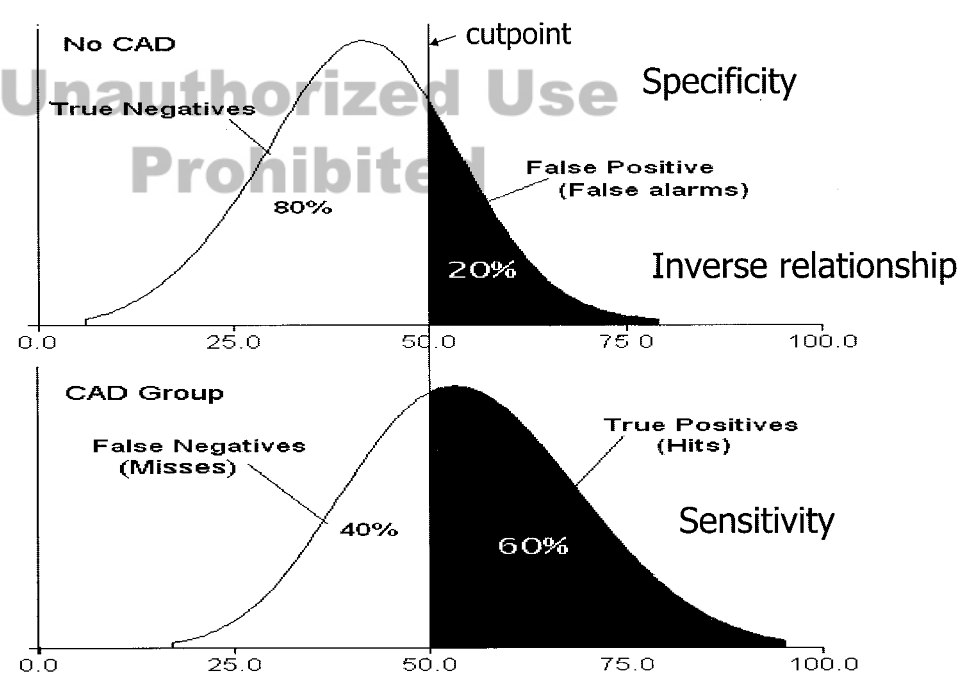

FIGURE 2—Range of characteristics plots of the simple treadmill score showing how different cut points can be chosen according to the specific use of the test.

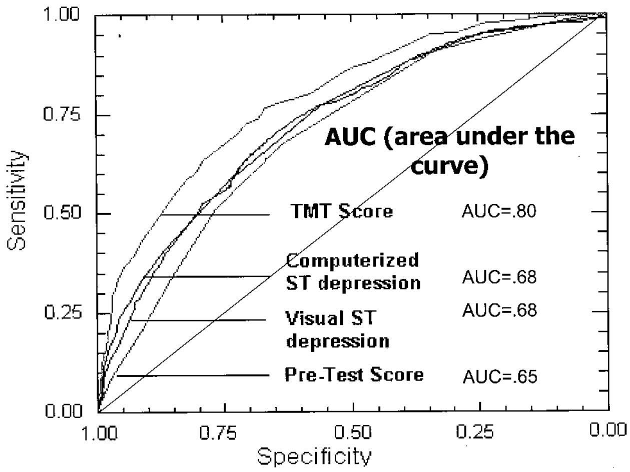

FIGURE 3—Range of characteristics plots comparing the discriminating power of a pretest score, ST measurements, and a simple treadmill score.

tivity is required. Figure 3 also illustrates a comparison of the diagnostic characteristics of a pretest clinical score, ST segment analysis alone, and a treadmill score. The four curves allow for a comparison of the diagnostic value of these techniques. The treadmill test clearly adds to the discriminatory value of clinical data. Surprisingly, attempts to use a computer to improve upon visual ST segment analysis of the exercise test have failed to improve its diagnostic value (11), though a score is clearly an improvement over ST analysis alone (44).

Even though a score demonstrating an ability to separate those with from those without disease functions in another group of patients (i.e., it is portable), it still needs to be confirmed that the score's calibration is the same. That is, does the value of the score for 70% probability of disease as determined in one group, for instance, still represent 70% instead of a different probability in another group?

Pretest Scores

The exercise ECG is the recommended diagnostic test for patients with an intermediate probability for CAD. In the ACC Exercise Testing Guidelines (13), the classification of pretest probability is enabled through a table considering age, gender, and chest pain characteristics using the Diamond-Forrester tabular method (4,5,15) (Table 2). The intermediate pretest probability category was assigned a Class I indication, whereas the low and high pretest probabilities were assigned Class IIb indications for exercise testing. Morise et al. (31) proposed a pretest score for categorizing patients with suspected CAD and normal resting electrocardiograms that possibly is superior to the method advocated by the guidelines (32). We have validated this score in a large sample of male veterans (37). It can be calculated using Figure 4.

EXERCISE TEST DIAGNOSTIC SCORES

Multi-Variate Prediction Equations/Meta-Analysis

Since the seminal work of Ellestad and colleagues (7) demonstrated that combining other clinical and exercise parameters along with the ST responses could improve the accuracy of the test, many clinical investigators have published studies proposing multivariable equations to enhance the accuracy of the standard exercise test. We reviewed 24 studies that used multivariate techniques to develop scores to predict presence of any angiographic disease from clinical and exercise test data (44). In Table 3, the number of studies that selected each given variable to be a significant predictor of CAD in the multivariate model is the numerator, and the total number of studies that considered that variable is the denominator. The variables selected in more than half of the studies are in bold.

Management Strategy Using Scores

Scores can also provide a management strategy for patients with possible CAD. This is done by placing patients into three categories of risk rather than just dichotomizing them as positive or negative (14). Low-risk patients have an excellent prognosis and may be risk-stratified by the treadmill test. This patient cohort can be managed safely with watchful follow-up as well as with symptomatic medical therapy without further testing. High-risk patients should be considered candidates for more aggressive management that may include cardiac catheterization. In patients with an intermediate-probability treadmill score, myocardial perfusion imaging or other tests appear to be of value for further risk stratification (Table 4) (2,13,17,23,24).

Consensus of Scores

In an attempt to make the scores function in different populations as well as they did in the population for which they were derived, a consensus approach was considered (6). Knowing that NASA calculates spacecraft trajectories

TABLE 2. Pretest probability of coronary disease by symptoms, gender, and age.

Age	Sex	Typical/Definite Angina Pectoris	Atypical/Probable Angina Pectoris	Non-Anginal Chest Pain	Asymptomatic
30–39	Men	Intermediate	Intermediate	Low (<10%)	Very low (<5%)
	Women	Intermediate	Very Low (<5%)	Very low	Very low
40–49	Men	High	Intermediate	Intermediate	Low
	Women	Intermediate	Low	Very low	Very low
50–59	Men	High (>90%)	Intermediate	Intermediate	Low
	Women	Intermediate	Intermediate	Low	Very low
60–69	Men	High	Intermediate	Intermediate	Low
	Women	High	Intermediate	Intermediate	Low
High = >90%		Intermediate = 10–90%	Low = <10%	Very Low = <5%	

There is no data for patients younger than 30 or older than 69, but it can be assumed that coronary artery disease prevalence increases with age.

Variable	Circle response	Sum
Age	Men<40, Women<50 = 3	
	Men40-55, Wom50-65=6	
	Men>55, Women>65 =9	
Estrogen Status	Positive=-3	
	Negative=+3	
Diabetes	Yes = 2	
Obesity?	Yes = 1	
Family History?	Yes = 1	
Hypercholesterolemia?	Yes = 1	
HBP?	Yes = 1	
Smoking?	Yes = 1	
	Total Score	

Choose only one per group

<=8 low prob

9-15 = intermediate probability

>=16 high probability

FIGURE 4—Calculation of the clinical score for angiographic coronary disease.

using a number of equations and then uses those that agree, the same method was used for predicting CAD. The Detrano (3) and Morise (32) equations provided in the ACC Exercise Testing Guidelines (14) were used along with an equation derived from a VA population. A probability score was calculated for each patient using the VA equation and the other two validated equations. Thresholds were set for each equation such that if a patient was high probability in at least two of the three equations, the patient was considered high probability; similarly, if low in at least two of the equations, the patient was considered low risk. All others would be intermediate. Since the patients in the intermediate group would be sent for further testing and would eventually be correctly classified, the sensitivity of the consensus approach was 94%, and specificity was 92%. The percentage of correct diagnoses increased from 67% for the standard exercise ECG analysis and from 77% for multivariable predictive equations alone to greater than 90% correct diagnoses for the consensus approach. However, this

TABLE 3. Results from meta-analysis of studies with angiographic findings as the gold standard for any significant coronary disease.

Variables	Fraction and % of time a variable is selected as a significant predictor when the variable was considered	
Clinical variables		
Gender	**20/20**	**100%**
Chest pain symptoms	**17/18**	**94%**
Age	**19/27**	**70%**
Elevated cholesterol	**8/13**	**62%**
Diabetes mellitus	6/14	43%
Smoking history	4/12	33%
Abnormal resting ECG	4/17	24%
Hypertension	1/8	13%
Family history of CAD	0/7	0%
Exercise test variables		
ST segment slope	**14/22**	**64%**
ST segment depression	**17/28**	**61%**
Maximal heart rate	**16/28**	**57%**
Exercise capacity	11/24	46%
Exercise induced angina	11/26	42%
Double product	2/13	15%
Maximal systolic BP	1/12	8%

Variables in *bold* chosen more than half the time.

TABLE 4. Paradigm for the clinical reaction to the score-estimated, stratified probability for angiographic coronary disease.

Probability for clinically significant CAD	Clinical Reaction
Low probability	Patient reassured symptoms most likely not due to CAD
Intermediate probability	Require other tests, such a stress echo, nuclear, or angiography to clarify diagnosis; anti-anginal medications tried.
High probability	Anti-anginal treatment indicated; intervention clinically appropriate; angiography may be required

approach can only be practically applied utilizing a computer program (42).

"SIMPLIFIED" SCORE DERIVATION

Simplified scores derived from multivariable equations have been developed for pretest estimates of disease and for prognosis. They require physicians only to add points. To develop such a score, data from two VA Medical Centers were analyzed (37). All patients had coronary angiography within 3 months of the exercise treadmill test. The score derived was then validated in 476 males from another institution.

To decrease the complexity of the predictive equations, we compiled the variables chosen in logistic regression into a simple linear score. We first coded all variables with the same number of intervals so that the coefficients would be proportional. Then we coded the category with the larger value to be associated with higher probability of disease. For instance, if 5 is the chosen interval, dichotomous variables are 0 if not present and 5 if present, and continuous variables such as age and heart rate are coded in 5 categories by appropriate ranges. All codes would then be directly related to probability (i.e., a heart rate code of 5 would be a low heart rate while an age code of 5 would be the oldest individuals) and the smallest coefficient is associated with the least important variable. The coefficient of this least important variable was divided into the other coefficients. This makes the relative importance of the selected variables readily apparent. This approach results in a simple linear

Variable	Circle response	Sum
Maximal Heart Rate	Less than 100 bpm = 30	
	100 to 129 bpm = 24	
	130 to 159 bpm =18	
	160 to 189 bpm =12	
	190 to 220 bpm =6	
Exercise ST Depression	1-2mm =15	
	> 2mm =25	
Age	>55 yrs =20	
	40 to 55 yrs = 12	
Angina History	Definite/Typical = 5	
	Probable/atypical =3	
	Non-cardiac pain =1	
Hypercholesterolemia?	Yes=5	
Diabetes?	Yes=5	
Exercise test	Occurred =3	
induced Angina	Reason for stopping =5	
	Total Score:	

Males

Choose only one per group

<40=low prob

40-60= intermediate probability

>60=high probability

FIGURE 5—Calculation of the simple score for angiographic coronary disease in men.

Variable	Circle response	Sum
Maximal Heart Rate	Less than 100 bpm = 20	
	100 to 129 bpm = 16	
	130 to 159 bpm =12	
	160 to 189 bpm =8	
	190 to 220 bpm =4	
Exercise ST Depression	1-2mm =6	
	> 2mm =10	
Age	>65 yrs =25	
	50 to 65 yrs = 15	
Angina History	Definite/Typical = 10	
	Probable/atypical =6	
	Non-cardiac pain =2	
Smoking?	Yes=10	
Diabetes?	Yes=10	
Exercise test	Occurred =9	
induced Angina	Reason for stopping =15	
Estrogen Status	Positive=-5, Negative=5	
	Total Score	

Women

Choose only one per group

<37=low prob

37-57= intermediate probability

>57=high probability

FIGURE 6—Calculation of the simple score for angiographic coronary disease in women.

score in which the health care provider merely compiles the variables in the score, multiples by the appropriate number and then adds up the products. Calculation of the simple score can be done using Figure 5. This diagnostic score did not perform well in women (AUC $<$ 0.65), and so a separate score was developed for women (see Fig. 6).

Some test results are dichotomous (normal, abnormal; positive, negative) rather than continuous like a score. Examples of these are perfusion defects, wall motion abnormalities, and coronary calcification. Any score can be dealt with as a dichotomous variable by choosing a cut point. For comparing dichotomous test results, the calculated predictive accuracy (percent of total true calls, positive and negative) can be used to compare the diagnostic characteristics of tests. However, predictive accuracy is affected by disease prevalence while ROC curves are not. Therefore, to compare tests, they must be evaluated in populations with roughly the same prevalence of disease. An advantage of predictive accuracy is that it provides an estimate of the difference in number of patients correctly classified by the test out of 100 tested. Table 5 summarizes the meta-analyses of the major diagnostic tests currently available with their predictive accuracies. It can be seen that only about 5 more patients per 100 tested are correctly diagnosed using the more expensive imaging tests compared with the standard exercise test using scores. Furthermore, using a score strategy to selectively determine who is referred for the more expensive tests has the greatest predictive accuracy.

TABLE 6. The common mistakes made by researchers attempting to determine the prognostic characteristics of a test.

- Limited challenge and workup bias
- Incomplete follow-up
- Failure to censor
- Use of misleading endpoints

SURVIVAL ANALYSIS

Different statistical techniques should be used to develop prognostic scores and predict outcomes. The key features of survival analysis are consideration of time to event and censoring. A Cox proportional hazards model should be used to select significant variables and to determine the effect of a given independent variable on time to death or event. To develop a simple score, each regression coefficient should be divided by the smallest coefficient in the score (17). Table 6 lists the most common mistakes made in prognostic studies that can invalidate their results in addition to those listed in Table 1.

End Points and Censoring

The relative importance of the ischemic variables can be minimized by not censoring on interventions for ischemia (i.e., removal of intervened patients from observation when the intervention occurs in follow up) and the consideration of all-cause mortality instead of cardiovascular mortality. This may also explain why the ischemic variables included in the Duke score that clearly had diagnostic power (41) do not predict all-cause mortality. While all-cause mortality has advantages over cardiovascular mortality as an endpoint (25), the Duke score was generated using the endpoints of infarction and cardiovascular death (10,28). The use of interventions as endpoints falsely strengthens the association of ischemic variables with endpoints since the ischemic responses clinically result in the intervention being performed. While some investigators have justified their use by requiring a time period to expire after the test before using the intervention/procedure as an endpoint, this still influences the associations between test responses and endpoints.

TABLE 5. The diagnostic characteristics of the major tests currently available.

Testing Method	Studies (*N*)	Total Patients (*N*)	Sensitivity	Specificity	Predictive Accuracy
Standard exercise test	147	24,047	68%	77%	73%
Exercise test scores	24	11,788			80%
Score strategy	2	>1000	85%	92%	88%
Thallium scintigraphy	59	6,038	85%	85%	85%
SPECT	16+14	5,272	88%	72%	80%
Adenosine SPECT	10+4	2,137	89%	80%	85%
Exercise ECHO	58	5,000	84%	75%	80%
Dobutamine ECHO	5	<1000	88%	84%	86%
Dobutamine scintigraphy	20	1014	88%	74%	81%
Electron beam tomography (EBCT)	16	3,683	60%	70%	65%

By subtracting the estimated predictive accuracies, the difference in the number of patients correctly classified or diagnosed out of 100 tested can be determined. For instance, 7 more patients are correctly classified out of 100 if scores are used compared with classifying patients as diseased or not diseased using only ST analysis. The use of a score strategy to determine when other tests should be ordered provides better results than only using any of the more expensive tests. Remember that sensitivity and specificity are easily altered by changing the test measurement cut point, and that all of the studies that are reported in these meta-analyses have been greatly influenced by workup bias, which raises sensitivity and lowers specificity.

TABLE 7. Frequency of clinical and exercise test variables chosen as significantly and independently associated with time until death/event in the nine major prognostic studies.

Variable	Number of Studies
Clinical	
Age	2
CHF	2
MI by history or Q waves	1
Resting ST depression	1
Exercise responses	
Exercise capacity (METs)	7
Angina	5
ST depression	4
Maximal heart rate	3
Maximal SBP	2
ST elevation	1
PVC's	1
Maximal double product	1

EXERCISE TEST PROGNOSTIC SCORES

Nine studies have incorporated multiple exercise variables into prognostic scores to extract the maximum information available and to allow the clinician to summarize the most important information from the test without using complex regression formulas. Table 7 lists the number of times the major prognostic variables were chosen as significantly and independently predictive of time to death out of the times they were considered in the published prognostic studies (9,44). The most accepted of these is the Duke Treadmill Score since it can be used both for prognosis (28) and diagnosis (41). This score has been validated in other populations, including women and when the resting ECG exhibits ST depression (1,8,10,19,24,29,33). The Duke Treadmill Score is calculated as follows:

Exercise time in the Bruce protocol – (5*ST depression) – (4* treadmill angina index)

The treadmill angina index is 1 if angina occurred and 2 if it was the reason for stopping. A nomogram using this score has been presented which facilitates estimation of prognosis (28).

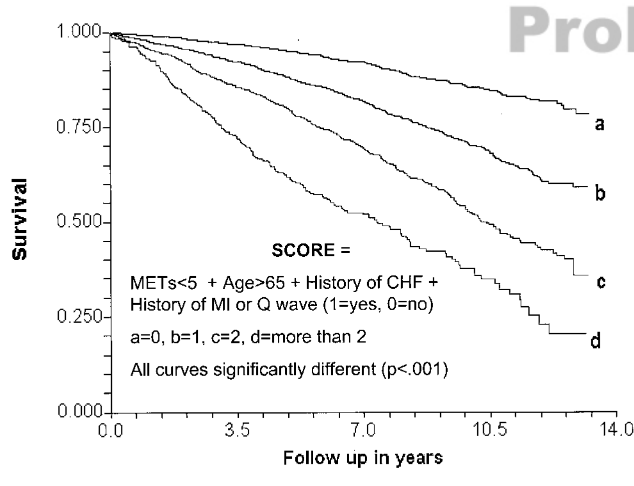

FIGURE 7—Kaplan-Meier Survival curves for the "all-comers" prognostic score.

Prognosis in "All-Comers" to the Exercise Lab

Previous prognostic studies focused on specific subsets of patients. We decided to analyze all patients referred for evaluation at our exercise lab between 1987 and 2000 in order to determine the prevalence of exercise test abnormalities (36). Cox hazard function chose the following variables in rank order as independently and significantly associated with time to death: metabolic equivalents less than 5, age greater than 65, history of congestive heart failure, and history of myocardial infarction. Metabolic equivalents were defined as multiples of the resting metabolic rate estimated from peak treadmill speed and grade. A score based on simply adding these variables classified patients into low, medium, and high-risk groups. Figure 7 illustrates marked differences in survival between each risk group.

COMPARING SCORES AND PHYSICIANS

Though scores based on exercise testing data have been advocated for years, only three previous studies have compared them to physician estimates of disease. Detrano and colleagues performed one of the first such studies (3). They derived a score for estimating probabilities of significant and severe coronary disease and then validated and compared it with the assessments of cardiologists. The score performed at least as well as the clinicians when the latter knew the identity of the patients. The clinicians were more accurate when they did not know the identity of the subjects but worked from tabulated objective data. Hlatky and colleagues validated two scores by comparing their diagnostic accuracy to that of cardiologists (17). The scores outperformed these cardiologists. A third study considered scores for prognosis (rather than diagnosis) with 100 patients sent to five senior cardiologists at one center (26). Again the scores outperformed these cardiologists.

We performed a study that was larger and included different groups of physicians, validating these earlier studies (27). Five hundred ninety-nine consecutive male patients without prior MI and with a mean age of 59 ± 11 yr were considered for this analysis. The clinical/treadmill test reports were sent to expert cardiologists and to two other groups including randomly selected cardiologists and internists who classified them as high, low, or intermediate probability of disease in addition to estimating a numerical probability from 0 to 100%. Forty-five expert cardiologists returned estimates on 336 patients, 37 randomly chosen practicing cardiologists returned estimates on 129 patients, 29 randomly chosen practicing internists returned estimates on 109 patients, 13 academic cardiologists returned estimates on 102 patients, and 27 academic internists returned estimates on 174 patients. When probability estimates were compared, the scores were superior to all the physician groups. In a subsequent analysis, we found the scores predicted prognosis as well or better than physicians.

CONCLUSIONS

Physicians should not reduce their diagnostic assessments to blindly using and memorizing prediction rules (18,21). Scores should complement, not replace, the knowledge and experience the clinician draws from when making clinical decisions (38). In spite of the methodological limitations of the available studies, scores help to facilitate evidence-based clinical decisions. Statistical approaches cannot make counter-intuitive leaps of tangential thinking, but they excel at that which humans do not: considering vast quantities of information, then categorizing and analyzing it without bias and developing scores that can help make diagnoses. Making use of the statistics described herein gives clinicians a powerful second opinion and allows them to concentrate on what the computer can never do: assess and treat patients as individuals. In particular, scores make available the experience of the specialist clinician to generalists. Generalists have to cover a wide range of specialties and they cannot be equally up to date in each. Multivariate scores can, in certain cases, equal the diagnostic reasoning of specialist physicians. Making these 'opinions' available to the generalist would allow resources to be concentrated on those who need them the most. Scores can help diagnose, thereby avoiding expensive, unnecessary invasive investigations and their associated risk and help with prediction of prognosis, allowing optimal use of secondary prevention measures.

Address correspondence to: Victor Froelicher, M.D., Cardiology Division (111C), VA Palo Alto Health Care System, 3801 Miranda Ave., Palo Alto, CA 94304; Email: vicmd@aol.com; internet: www.Cardiology.org.

REFERENCES

1. Alexander K., L. Shaw, E. Delong, et al. Value of exercise treadmill testing in women. *J. Am. Coll. Cardiol.* 32:1657–1664, 1998.
2. Chatziioannou S. N., W. H. Moore, P. V. Ford, et al. Prognostic value of myocardial perfusion imaging in patients with high exercise tolerance. *Circulation* 99:867–872, 1999.
3. Detrano R., M. Bobbio, H. Olson, et al. Computer probability estimates of angiographic coronary artery disease: transportability and comparison with cardiologists' estimates. *Comput. Biomed. Res.* 25:468–485, 1992.
4. Diamond G, and J. Forrester. Analysis of probability as an aid in the clinical diagnosis of CAD. *N. Engl. J. Med.* 300:1350–1358, 1979.
5. Do D., J. West, A. Morise, E. Atwood, and V. Froelicher. A consensus approach to diagnosing coronary artery disease based on clinical and exercise test data. *Chest* 11:1742–1749, 1997.
6. Ellestad M. H., S. Savitz, D. Bergdall, and J. Teske. The false positive stress test. Multivariate analysis of 215 subjects with hemodynamic, angiographic and clinical data. *Am J Cardiol* 40: 681–687, 1977.
7. Elveback L. R., D. C. Connolly, and L. T. Kurland. Coronary heart disease in residents of Rochester, Minnesota. II. Mortality, incidence, and survivorship, 1950-1975. *Mayo Clin. Proc.* 56: 665–672, 1981.
8. Fearon W., D. Lee, and V. Froelicher. The effect of resting ST segment depression on the diagnostic characteristics of the exercise treadmill test. *J. Am. Coll. Cardiol.* 35:1206–1211, 2000.
9. Froelicher V. F. and J. Myers. *Exercise and the Heart,* 4th Ed. Philadelphia, PA: W.B. Saunders, 2000.
10. Froelicher V., K. Morow, M. Brown, E. Atwood, and C. Morris. Prediction of atherosclerotic cardiovascular death in men using a prognostic score. *Am. J. Cardiol* 73:133–138, 1994.
11. Froelicher V. F., K. G. Lehmann, R. Thomas, et al. The electrocardiographic exercise test in a population with reduced workup bias: diagnostic performance, computerized interpretation, and multivariable prediction. Veterans Affairs Cooperative Study in Health Services #016 (QUEXTA) Study Group. Quantitative Exercise Testing and Angiography. *Ann. Intern. Med.* 128:965–974, 1998.
12. Garber A. M. and N.A. Solomon. Cost-effectiveness of alternative test strategies for the diagnosis of coronary artery disease. *Ann. Intern. Med.* 130:719–728, 1999.
13. Gibbons R., K. Chatterjee, J. Daley, et al. ACC/AHA/ACP-ASIM guidelines for the management of patients with chronic stable angina: executive summary and recommendations. A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Management of Patients with Chronic Stable Angina). *Circulation* 99:2829–2848, 1999.
14. Gibbons R. J., G. J. Balady, J. W. Beasley, et al. ACC/AHA guidelines for exercise testing: executive summary. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* 96:260–311, 1997.
15. Guyatt G. H. Readers' guide for articles evaluating diagnostic tests: What ACP Journal Club does for you and what you must do yourself. *ACP Journal Club* 115:A-16, 1991.
16. Hachamovitch R., D. S. Berman, H. Kiat, et al. Exercise myocardial perfusion SPECT in patients without known coronary artery disease: incremental prognostic value and use in risk stratification. *Circulation* 93:905–914, 1996.
17. Hlatky M., E. Botvinick, and B. Brundage. Diagnostic accuracy of cardiologists compared with probability calculations using Bayes' rule. *Am J Cardiol* 49:1927–1931, 1982.
18. Hlatky M. A., R. M. Califf, F. E. Harrell, K. L. Lee, D. B. Mark, L. H. Muhlbaier, and D. B. Pryor. Clinical judgment and therapeutic decision making. *J. Am. Coll. Cardiol.* 15:1–14, 1990.
19. Iscandrian A. S., M. Ghods, H. Helfeld, et al. The treadmill exercise score revised: coronary arteriographic and thallium perfusion correlates. *Am. Heart J.* 124:1581–1586, 1992.
20. Deleted in proofs.
21. Knottnerus A. Diagnostic prediction rules: principles, requirements and pitfalls. *Primary Care.* 22:341–363, 1995.
22. Kotler T. S. and G. A. Diamond. Exercise thallium-201 scintigraphy in the diagnosis and prognosis of coronary artery disease. *Ann. Intern. Med.* 113:684–702, 1990.
23. Kuntz K. M., K. E. Fleischmann, M. Hunink, and P. S. Doulas. Cost-effectiveness of diagnostic strategies for patients with chest pain. *Ann. Intern. Med.* 130:709–718, 1999.
24. Kwok J. M., Miller T. D., Christian T. F., Hodge D. O., Gibbons R. J. Prognostic value of a treadmill exercise score in symptomatic patients with nonspecific ST-T abnormalities on resting ECG. *JAMA* 282:1047–1053, 1999.
25. Lauer, M. S., E. H. Blackstone, J. B. Young, and E. J. Topol. Cause of death in clinical research: time for a reassessment? *J. Am. Coll. Cardiol.* 34:618–620, 1999.
26. Lee K. L., D. B. Pryor, F. E. Harrell, Jr, et al. Predicting outcome in coronary disease: statistical models versus expert clinicians. *Am. J. Med.* 80:553–560, 1986.
27. Lipinski M., V. Froelicher, B. Franklin, et al. Can physicians predict disease as well as physicians? *Arch Internal Medicine* 161:2239–2244, 2001.
28. Mark D. B., L. Shaw, F. E. Harrell Jr., et al. Prognostic value of a treadmill exercise score in outpatients with suspected coronary artery disease. *N. Engl. J. Med.* 325:849–853, 1991.
29. Morise A. and R. Duval. Comparison of three Bayesian methods to estimate posttest probability in patients undergoing exercise stress testing. *Am. J. Cardiol* 64:1117–1122, 1989.

30. MORISE A., J. HADDAD, and D. BECKNER. Development and validation of a clinical score to estimate the probability of coronary artery disease in men and women presenting with suspected coronary disease. *Am. J. Med.* 102:350–356, 1997.
31. MORISE A. Comparison of the Diamond-Forrester method and a new score to estimate the pretest probability of coronary disease before exercise testing. *Am. Heart J.* 138:740–745, 1999.
32. MORISE A. P., R. DETRANO, M. BOBBIO, and G. A. DIAMOND. Development and validation of a logistic regression-derived algorithm for estimating the incremental probability of coronary artery disease before and after exercise testing. *J. Am. Coll. Cardiol.* 20:1187–1196, 1992.
33. MORROW K., C. MORRIS, V. FROELICHER, A. HIDEG, et al. Prediction of cardiovascular death in men undergoing noninvasive evaluation for coronary artery disease. *Ann. Intern. Med.* 118:689–695, 1993.
34. PHILBRICK J. T., HOROWITZ, A. R. FEINSTEIN. Methodological problems of exercise testing for coronary artery disease: groups, analysis and bias. *Am. J. Cardiol.* 46:807–812, 1980.
35. POSES R. M., R. D. CEBUL, M. COLLINS, et al. The importance of disease prevalence in transporting clinical prediction rules. *Ann. Intern Med.* 105:586–591, 1986.
36. PRAKASH M., V. FROELICHER, J. MYERS, et al. Clinical and exercise test predictors of all-cause mortality: results from > 6,000 consecutive referred male patients. *Chest* 120:1003–1013, 2001.
37. RAXWAL V., K. SHETLER, A. MORISE, D. DO, J. MYERS, J. E. ATWOOD, and V. FROELICHER. Simple treadmill score to diagnose coronary disease. *Chest* 119:1933–1940, 2001.
38. REEDER G. S. and B. J. GERSH. Modern management of acute myocardial infarction. *Curr. Prob. Cardiol.* 21:590–662, 1996.
39. REID M., M. LACHS, and A. FEINSTEIN. Use of methodological standards in diagnostic test research. *JAMA* 274:645–651, 1995.
40. ROSAMOND W. D., L. E. CHAMBLESS, A. R. FOLSOM, et al. Trends in the incidence of myocardial infarction and in mortality due to coronary heart disease, 1987 to 1994. *N. Engl. J. Med.* 339:861–867, 1998.
41. SHAW L. J., E. D. PETERSON, L. K. SHAW, et al. Use of a prognostic treadmill score in identifying diagnostic coronary disease subgroups. *Circulation* 98:1622–1630, 1998.
42. SHIU P. and V. FROELICHER. An expert system for exercise testing utilizing consensus to predict coronary disease. *J. Inv Card* 2:21–26, 1998.
43. SWETS J. A., R. M. DAWES, and J. MONAHAN. Better decisions through science. *Sci. Am.* 283:82–87, 2000.
44. YAMADA H., D. DO, A. MORISE, J. E. Atwood, and V. FROELICHER. Review of studies using multivariable analysis of clinical and exercise test data to predict angiographic coronary artery disease. *Prog. Cardiovasc. Dis.* 39:457–481, 1997.